THESIS

ESTIMATING DIURNAL PATTERNS OF LAND SURFACE TEMPERATURE USING VISION TRANSFORMERS AND SATELLITE IMAGES

Submitted by

Srivarshini Ksheerasagar

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2026

Master's Committee:

Advisor: Dr. Sangmi Lee Pallickara
Co-Advisor: Dr. Shrideep Pallickara

Dr. Phuong D. Dao

Copyright by Srivarshini Ksheerasagar 2026

All Rights Reserved

ABSTRACT

ESTIMATING DIURNAL PATTERNS OF LAND SURFACE TEMPERATURE USING VISION TRANSFORMERS AND SATELLITE IMAGES

Diurnal cycles, the recurring 24–hour patterns produced by Earth's rotation shape a wide range of environmental processes including temperature variation, evapotranspiration, and soil thermal dynamics. Land surface temperature (LST), one of the 54 Essential Climate Variables defined by the Global Climate Observing System, serves as a central parameter in climatological, hydrological, agricultural, and ecological studies. However, obtaining complete diurnal LST patterns remains difficult. The sparse coverage of *in situ* stations, together with cloud contamination, environmental factors, sensor outages, and scan mismatches in satellite imagery, interrupt temporal continuity and leave large gaps in the record.

This study introduces DAYVIEW, a spatiotemporal deep learning framework designed to reconstruct full diurnal cycles of LST from a single satellite observation, regardless of acquisition time. The methodology draws on hourly products from the GOES–R satellite series over the contiguous United States and integrates ancillary information such as climatic zones and elevation. Built on a Vision Transformer (ViT) architecture with a Masked Autoencoder strategy, DAYVIEW directly addresses three core challenges: (1) estimating diurnal cycles from sparse observations, (2) incorporating environmental context to refine fluctuation modeling, and (3) extending predictions reliably across continental scales.

Empirical validation using remote sensing datasets demonstrates that DAYVIEW achieves high accuracy and strong robustness across diverse spatial and temporal conditions. Because the method is not limited to temperature alone, it can also be applied to other diurnal phenomena, such as solar–induced fluorescence, thus advancing environmental monitoring, climate analysis, and decision making at scale.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Sangmi Lee Pallickara for her guidance, encouragement, and patience throughout the course of this research. Their insight and expertise were invaluable in shaping this thesis, and her unwavering belief in my abilities gave me the motivation throughout the research process.

I am also thankful to my committee members, Dr. Shrideep Pallickara and Dr. Phuong D. Dao for their time, thoughtful feedback and constructive suggestions which improved the quality of this thesis.

I would like to thank my colleagues for their insightful discussions, and for creating a supportive and motivating environment.

I am grateful to my family for their love, support, and constant encouragement throughout this journey. Finally, I would like to thank my friends for their understanding and encouragement, which helped me stay motivated during challenging times.

This research was supported by the National Science Foundation (1931363, 2312319), the National Institute of Food and Agriculture (2024-67021-43840, 2025-77039-45531)), an NSF/NIFA Artificial Intelligence Institutes AI-LEAF (2023-03616) and a Clare Booth Luce Professorship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

A *diurnal* cycle refers to the recurring 24-hour pattern produced by the Earth's rotation on its axis. Many natural processes, including temperature, cloud formation, precipitation, evapotranspiration, and soil thermal dynamics, rise and fall within this daily rhythm. However, because these phenomena change quickly and at sub-daily scales, they remain difficult to capture in their full form.

Observational stations provide high-frequency measurements but they are geographically sparse. Their limited coverage prevents them from adequately representing spatial variability across broad regions. Satellites, by contrast, provide wide coverage and frequent data that are often publicly available. However, their potential is diminished by a persistent obstacle: cloud occlusions. When frequent images are obscured, the continuity of records is broken, leaving gaps in the very data needed to reconstruct diurnal behavior. A consequence is that achieving reliable and spatially complete observations across large geographical extents remains extremely challenging.

Within the wide range of diurnal processes, this thesis research focuses on land surface temperature. Land surface temperature (LST) is one of the 54 Essential Climate Variables identified by the Global Climate Observing System as fundamental for monitoring climate change [3]. LST governs evapotranspiration, influences soil moisture, and contributes directly to the Earth's radiation budget. Because LST responds directly to daily cycles of solar radiation, its fluctuations provide a sensitive measure of the Earth's surface energy balance.

Understanding these diurnal fluctuations is essential for the analysis of meteorological, climatological, hydrological, and ecological processes. An accurate model of diurnal land surface temperature can identify peak heat periods from even a handful of observations, and provide a useful guide for irrigation scheduling. It can also sharpen assessments of human heat-stress risk by pinpointing the hottest parts of the day, which daily averages often fail to capture.

The crux of this research is to design a methodology that reconstructs the complete daily fluctuation of land surface temperature from very limited observations distributed across wide spatial extents. To accomplish this, we employ a generative deep learning approach applied to NASA's GOES satellite imagery of the United States. Our model is designed to infer the entire hourly pattern of land surface temperature from a single observation, regardless of the time of day at which that measurement occurs. To strengthen the model's ability to capture the unique relations between diurnal patterns across diverse landscapes and climactic conditions, we also integrate two slow-changing geospatial features: climatic zone information from Köppen climatic zone data and elevation information from digital elevation model(DEM). These in particular were chosen as they provide slow-changing varying, geographical context for a rapidly changing variable like LST over space and time. This provides the model with information about the underlying climate conditions and topography to help the model infer these environmental conditions that control surface temperature improving reconstruction consistency across regions. This capability directly addresses the usability of datasets that provide only once-daily values and scales effectively.

Although our immediate focus is land surface temperature, we posit that our framework (co-denamed DAYVIEW) generalizes to any diurnally varying variable, such as solar-induced fluorescence and can therefore enhance subsequent analyses and the decision-making processes that depend upon them. More broadly, it demonstrates how masked generative learning can be harnessed for spatiotemporal Big Data domains where incomplete observations hinder downstream analysis and decision making. At its core, our methodology builds upon the Masked Autoencoder with Visual Transformer architecture [2], trained on hourly land surface temperature products from the GOES-R satellites, to achieve reconstruction of diurnal cycles with both accuracy and generalizability.

1.1 Research Questions

To address the challenges of reconstructing diurnal land surface temperature patterns at scale, our study is organized around 3 central research questions. RQ-1 concerns the fundamental fea-

sibility of inferring complete diurnal cycles from sparse or incomplete observations. RQ-2 explores whether incorporating ancillary variables can improve the fidelity of reconstructed cycles by grounding them in additional signals linked to land surface processes. RQ-3 expands the scope of inquiry to the continental level, where spatial heterogeneity and temporal variability introduce further complexity.

RQ-1: *How can we design a data-driven model that estimates the full diurnal cycle of land surface temperature when measurements are available only at limited temporal intervals, such as once daily?* RQ-1 also encompasses situations in which the data record is sparse because of high rates of missing values, which further constrain the accuracy of modeling efforts. Whereas most existing approaches attempt to reconstruct isolated snapshots, we ask whether it is possible instead to generate an entire 24-hour sequence of predictions, thereby capturing the continuous pattern of fluctuation rather than a single moment in time.

RQ-2: *How can we integrate ancillary data that are closely related to land surface temperature in order to improve the reconstruction of its diurnal cycle?* RQ-2 addresses the potential of a deep learning framework to leverage additional variables that share physical or environmental connections with the target phenomenon to enhance the model's ability to capture fluctuation patterns with greater fidelity.

RQ-3: *How can we extend the estimation of diurnal land surface temperature cycles to large spatial extents?* RQ-3 entails not only capturing the temporal dynamics but also representing the variability that arises across heterogeneous landscapes. The goal is to to ensure that our methodology remains both accurate and computationally tractable when applied to broader geospatial domains.

1.2 Overview of Methodology

The DAYVIEW framework is designed to estimate hourly fluctuations in land surface temperature across large spatial scales when only partial satellite observations are available. Using imagery from the GOES-R geostationary satellites, the model predicts the complete sequence of 24 hourly

temperature maps for a given day from a single observation. This design addresses the pervasive problem of missing data caused by cloud occlusion and limited observation schedules, allowing reconstruction of diurnal patterns that would otherwise remain incomplete. To prepare the data for modeling, full-extent images are partitioned into smaller tiles through quadtree hashing and further divided into patches. Each patch is supplemented with ancillary features, including elevation and Köppen climate zones, to provide spatially aligned multi-modal inputs.

The architecture of DAYVIEW extends a Vision Transformer (ViT) pretrained with a Masked Autoencoder strategy. During training, portions of the data are intentionally withheld, and the model learns to infer the missing values from the visible ones. This process allows the encoder to capture both local and global spatial dependencies even when observations are incomplete. Unlike original ViT, we combine the regression multihead that generates 24 hourly timesteps of LST images. Multi-modal information is fused within the encoder, ensuring that both environmental conditions and spatial context inform the learned representations.

Geospatial and temporal encodings are incorporated to further improve performance. Latitude and longitude are projected into continuous sine and cosine functions, preserving relative distances and ensuring consistent representation across space. Similarly, hour of day and day of year are transformed into cyclical values that allow the model to learn diurnal and seasonal patterns without introducing artificial discontinuities. Together, these encodings provide the framework with the context needed to generate accurate and stable reconstructions of hourly land surface temperature across diverse landscapes and climates.

By uniting high-frequency satellite observations, multi-modal ancillary data, and spatiotemporal encodings within a generative deep learning architecture, DAYVIEW targets designing a scalable approach to modeling diurnal processes at large spatial scales. The framework is not confined to land surface temperature alone but can be extended to other phenomena with daily cycles, offering a generalizable methodology for extracting high-resolution temporal dynamics from sparse and imperfect Earth observation datasets.

1.3 Main Contributions

Here, we introduce DAYVIEW, a spatiotemporal deep learning framework that predicts the full diurnal pattern of land surface temperature across the Contiguous United States (CONUS). Our specific contributions include the following:

- A novel spatiotemporal transformer architecture that estimates hourly land surface temperature for an entire day from a single observation, regardless of the time of acquisition, with high predictive accuracy.
- We demonstrate that the framework maintains consistent performance across both spatial and temporal variability, achieving robust results at the CONUS scale.
- We provide extensive empirical validation, including evaluations against independent remote sensing datasets and other existing models. These comparisons confirm the superior performance of DAYVIEW relative to existing approaches and highlight its strong agreement with observed ground truth.

Together, these contributions enable DAYVIEW to scale effectively while reconstructing diurnal processes from sparse and imperfect observations. At the same time, the framework remains generalizable to large and heterogeneous geospatial domains.

Translational Impact: By reconstructing complete diurnal temperature cycles from sparse satellite observations, DAYVIEW can provide actionable inputs for irrigation scheduling, heat-stress monitoring, and ecological forecasting. Its scalability across CONUS demonstrates its potential to inform decision making in agriculture, public health, and environmental management where reliable high-frequency data remain scarce.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. In Chapter 2, we include a background and related work discussion. Chapter 3 describes our methodology. We profile several aspects of our work in Chapter 4. Finally, we outline our conclusions and future work in Chapter 5.

Chapter 2

Background and Related Works

2.1 Early Attempts of Diurnal Models

Early attemps for diurnal modelling consisted of basic math to estimate how temperatures rise during the day and fall during the night. Parton and Logan's work [**?**] introduces a dual-phase representation that combined a truncated sinusoidal daytime curve with an exponential nocturnal delay. This made it possible to accurately reconstruct hourly data from only daily extrema. Salinger and Griffiths put forward a beta-distribution–based framework in "A Model for the Diurnal Variation of Temperature" [4] around the same time. This framework normalizes time and temperature to allow for flexible descriptions of climatological variability. In "Evaluation of Models for Predicting Diurnal Variation in Air Temperature" [5], Reicosky et al. thoroughly tested these empirical methods. They found that the accuracy of the models depended a lot on the amount of cloud cover and the climate in the area. These early semi-emphirical and stochastic models although effective were limited in their ability to generalize across diverse land-surface types and atmospheric conditions.

2.2 Satellite-Driven Diurnal Modeling

The availability of geostationary and polar-orbiting satellite observations led to a transition in diurnal modeling towards physics-based and satellite-driven methodologies. Schädlich et al. and Göttsche and Olesen enhanced energy-balance formulations utilizing MSG–SEVIRI data in [**?**] and [6], thereby diminishing noise and augmenting temporal continuity in geostationary LST retrievals. Jin and Dickinson introduced an algorithmic interpolation method for generating cloud-free LST [7]. This method combines AVHRR observations with modeled spatial–temporal patterns. Pinker and Pinker devised an empirical GOES-based DTC reconstruction method [8] whereas Inamdar et al. introduced a versatile four-parameter model for MODIS [9]. These satellite driven approaches greatly improve estimation of surface temperature by increasing the number of samples taken over

time and making it less dependent on sparse in-situ data. However, there were still problems with cloud cover, uneven land surfaces, and incomplete sampling. This led to the development of modern machine-learning and data-fusion frameworks.

2.3 Diurnal Models with Deep Learning Approaches

Deep learning has changed how we model the diurnal patterns by making it possible to use data-driven approximations of complicated surface-atmosphere interactions that traditional parametric models have trouble with. In the beginning, most applications used convolutional neural networks (CNNs) and recurrent architectures to fill in missing land surface temperature (LST) data from satellite time series. Duffy et al. showed that CNN-based multisensor fusion can give LST estimates that are almost real-time and are less affected by clouds and changes in view angle [10]. Their framework demonstrated that deep feature extractors surpass traditional interpolation by utilizing complementary temporal sampling from various satellite platforms.

Later work built on these ideas by adding ensemble learning and spatiotemporal fusion techniques. Liu et al. presented DELAG, a deep ensemble model explicitly engineered for the reconstruction of high-resolution land surface temperature (LST) in the presence of cloud cover [11]. This method uses parallel learns to make predictions more stable and less uncertain. It also makes use of the multi-scale features to capture Diurnal curves across different types of Land Surfaces. This approach persisted with the emergence of sophisticated fusion frameworks, represented by the high-resolution hourly Land Surface Temperature (LST) product developed by Su et al. [12], which uses deep learning to integrate multi source observations and explicitly simulates diurnal thermal behavior at 30m spatial resolution. Similarly, Kustura et al. used multi-sensor spatiotemporal fusion networks to get high-resolution temperature values that keep fine-scale spatial details and temporal coherence [13]. Han et al. expanded upon these concepts and created a time-continuous LST product by integrating deep learning with multi-source remote-sensing databases, demonstrating the ability of their model to bridge temporal gaps while preserving consistency [14]. Despite these advances, deep-learning models still face several limitations as most supervised models depend

on large volumes of temporally complete data which is difficult to obtain in cloudy or seasonally variable regions and land cover type.

2.3.1 Self-supervised Machine Learning

While supervised deep-learning approaches for surface temperature reconstruction have yielded good results, they require could-free training datasets which is difficult to obtain in real-world settings. Self-supervised learning (SSL) offers a compelling alternative by enabling models to learn rich spatio-temporal representations from vast archives of unlabeled or partially observed satellite imagery, reducing dependence on dense LST labels.

Recent work by Goh et al. showed that masked autoencoder (MAE) can reconstruction high resolution sea surface temperature (SST) even under heavy cloud coverage with high fidelity [15]. Their results suggest MAE-based models can implicitly learn physical structures and continuity in temperature fields, even when large proportions of data are missing which is their key advantage for diurnal cycle modeling under cloud cover.

Building on this, scale-aware SSL frameworks such as *Scale-MAE* [16] have been proposed for geospatial data: by explicitly encoding information across spatial scales, such methods improve generalization across different land cover types and resolutions. Similarly, *Fus-MAE* [17] demonstrates that masked autoencoders combined with cross-attention fusion can handle multi-sensor data (e.g., SAR + optical), bridging domain gaps and enabling flexible data fusion in a self-supervised manner. These advances show that SSL is not only feasible but potentially powerful for heterogeneous remote-sensing tasks.

Applied results in land surface temperature reconstruction by Wu et al. [18] — underlined the value of robust reconstruction techniques given the frequent gaps and noise in satellite-derived LST. By leveraging SSL, latent representations learned from large, unlabeled satellite archives can be fine-tunes or decoded to produce diurnal LST generated values with higher robustness to cloud cover, variable sampling and land cover heterogeneity. Given the scarcity of dense, global, hourly

LST 'ground truth', SSL provides a scalable path forward, improving generalization and reducing reliance on fully labeled datasets.

Thus, self-supervised machine Learning, particularly masked auto encoding and multi-sensor fusion forms a promising methodological foundation for our approach to diurnal LST modeling.

2.3.2 Vision Transformers for Diurnal LST Modeling

Vision Transformers(ViT) is a neural network architecture that consists of a transformer block that was initially made for Natural Language processing. ViT don't use convolution layers to process images instead, they break input images into series of patches that don't overlap. These patches are later flatten and each patch is then projected into a fixed-dimensional embedding space. After that, these patch embeddings are combined with positional encodings to keep spatial information. They go through many layers of multi-head self-attention and feed-forward networks [1]. The self-attention mechanism lets the model weigh the relationships between all the patches in real time. This lets it capture both local and global spatial dependencies across the whole image. This is different from traditional convolutional neural networks (CNNs), which use fixed local receptive fields and hierarchical feature aggregation. This makes it hard for them to model long-range interactions well.

Vision Transformers (ViTs) are particularly effective for analyzing satellite imagery, which frequently exhibits complex spatial patterns because of variations in land cover, topography, and atmospheric influences. By dividing images into patches and employing global self-attention, ViTs are able to simultaneously capture fine-scale local details and broad spatial relationships, such as regional temperature gradients, making them well-suited for tasks that require understanding both local and large-scale structures. They do this by representing images as sequences of patches and using global self-attention mechanisms [19–22]. This feature is especially useful for recreating daily temperature cycles, where the way different landscapes interact with each other affects how the temperature changes during the day. So, ViTs are a flexible and powerful way to model LST data from satellites.

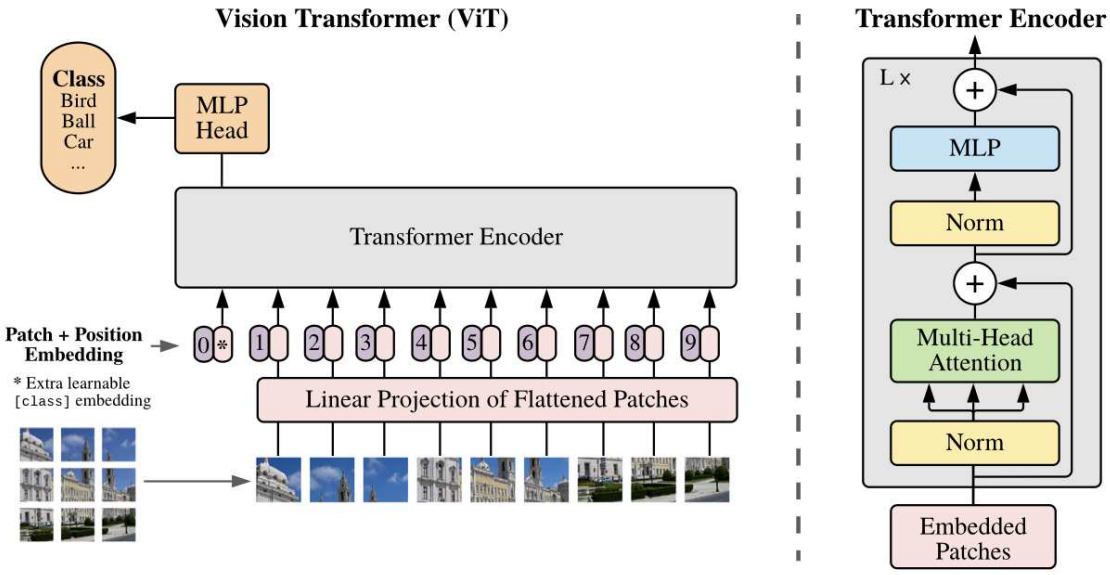

Figure 2.1: Vision Transformer Architecture [1]

2.3.3 Masked auto-encoding Vision Transformers

Masked Autoencoding Vision Transformers (MAE-ViTs) build on the Vision Transformer framework by using self-supervised pretraining. In this method, a large part of the input image patches are randomly hidden, and the model learns to fill in the gaps. [2]. This makes the network learn strong latent representations that include both local structures and global context, even when there is no labeled data. This kind of pretraining is especially helpful for satellite images because high-resolution, fully labeled datasets are often hard to find.

MAE-ViTs can model daily temperature cycles in a self-supervised way, which is useful for reconstructing land surface temperature (LST). The model can generalize across different types of land cover, topography, and weather conditions by learning from satellite images that are only partially visible. This effectively fills in gaps caused by clouds or irregular revisit intervals [15]. As a result, MAE-ViTs offer a scalable and adaptable framework for high-resolution spatio-temporal modeling of land surface temperature (LST) derived from satellite observations.

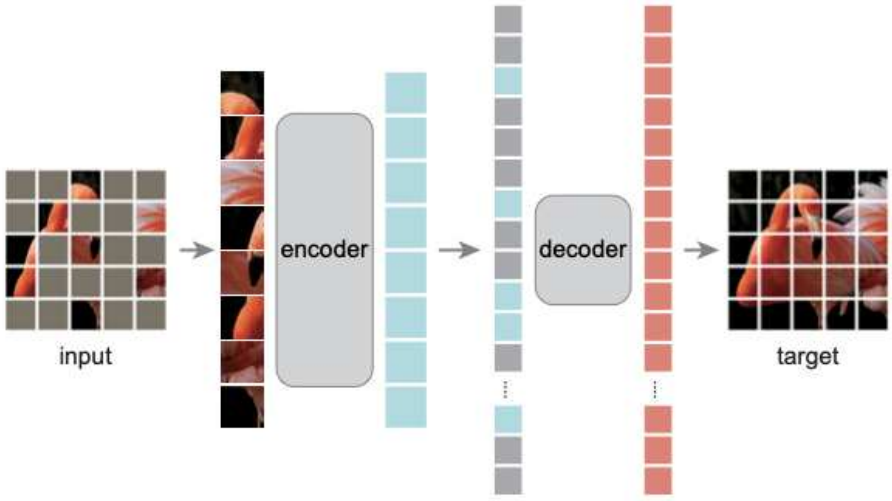

Figure 2.2: Masked Auto-encoder [2]

2.3.4 Science guided learning

Integrating scientific knowledge into deep learning models are based on physics-informed learning or knowledge-guided machine learning. These approaches provide improved physical realism and fidelity to phenomena [23]. KGML methods have been used to incorporate support for physical constraints grounded in soil hydrology, including the van Genuchten water retention equations and models (i.e., Richards' Equation) of hydraulic conductivity [24, 25]; vegetation indices [26]; evapotranspiration [27]; preservation of graph properties such as betweenness centrality [28]; masking cloud occlusions in satellite imagery [29]; accounting for human perceptual limits during visualization [30]; masked autoencoders [31, 32] and accounting for correlations between soil spectroscopic properties [33, 34]. Knowledge distillation approaches have also been explored in the context of physical phenomena [35]. Our methodology is complementary to these approaches.

Chapter 3

Methodology

Estimating fluctuations in land surface temperature across large spatial domains presents a significant challenge due to the sparsity of available observations. DAYVIEW addresses this challenge by using geostationary satellite imagery to capture hourly variations in temperature across the continental United States. Although these data provide broad coverage and high temporal resolution, they are often affected by extensive gaps caused by cloud occlusion. To overcome this limitation, DAYVIEW is designed to reconstruct the full daily cycle of hourly temperature maps from a single available observation, regardless of when that measurement occurs.

Our framework builds on the Vision Transformer (ViT) architecture and incorporates Masked Autoencoder pretraining. The transformer's ability to model image sequences aligns closely with the objective of generating a continuous series of hourly maps. However the conventional ViT–MAE approach is insufficient by itself. It does not adequately account for geospatial factors that shape diurnal temperature dynamics, including elevation, climate conditions, and time zones. In addition, the very high rate of missing data in the satellite product complicates the direct application of the original masking strategy. To account for these factors, DAYVIEW adapts and extends these techniques to meet the particular requirements of large-scale geospatial prediction.

In the sections that follow, we describe the key elements of our methodology. We begin with the preparation of the input dataset in Section 3.1. We then describe the overall model architecture in Section 3.2, followed by the design of the encoder and regression head in Section 3.3. Finally, we explain how spatiotemporal information is encoded to improve the accuracy and stability of the predictions in Section 3.4.

3.1 Data Sources, Study Regional, and Data Wrangling

This subsection describes the geographic domain of the study, the data sources employed, and the steps taken to prepare them for modeling. Because DAYVIEW requires inputs that are

both spatially and temporally consistent we targeted aligning satellite observations with ancillary datasets and also on addressing missing values in satellite imagery caused by cloud contamination. The following discussion proceeds from the description of data sources, through the study area and tiling strategy, to the methods used to produce spatially aligned inputs for model training and evaluation.

DAYVIEW is designed to estimate hourly maps of land surface temperature across the entire Contiguous United States (CONUS). Several satellite systems including Landsat, VIIRS, and the GOES series produce temperature maps derived from the thermal infrared band in combination with optical bands [36–38]. We rely on observations from the Geostationary Operational Environmental Satellite (GOES) series [8]. The satellite is in geostationary orbit at elevation approximately 22,300 miles, continuously observing the same regional footprint. This enables, high monitoring rate GOES-R series is the nation's most advanced fleet of geostationary weather satellites, designed to significantly improve the detection and observation of environmental phenomena that affect public safety, infrastructure, and economic interests across the Western Hemisphere. The ABI onboard GOES-R satellites provides imagery in 16 spectral bands including visible, near-infrared, and thermal infrared. ABI offers four times the spatial resolution and five times faster scanning which is a major improvement from the previous GOES generations which only had 5 bands [39]. This expanded spectral range allows improved discrimination of surface and atmospheric features and conditions which will provide a stronger context for variables like LST.

Using bands 14 (11.2 μm) and 15 (12.3 μm) in a split-window configuration, GOES-R produces land surface temperature estimates at a spatial resolution of 2 km and at a temporal resolution of five minutes, available for Full Disk, CONUS, and Mesoscale sectors. As ground truths for diurnal fluctuations, we employed a pre-generated hourly product [36] that provides land surface temperature values across CONUS. the ground truth data provides the verified reference information against which the model's predictions are compares to calculate accuracy. These full-domain images were partitioned into smaller tiles using quadtree hashing, which is a hierarchical representation of geographic information [40]. In this structure, space is recursively divided into quadrants,

each assigned a unique identifier within a tree data structure. Each node corresponds to a region and may be subdivided into four smaller subregions. After indexing the subregions, we cropped the images into 32×32 pixel patches and grouped them into sequences of 24 hourly images covering hours 0 through 23 at the same location. For model training, a single hourly image was randomly selected from each daily sequence and used as input, while the remaining images in the sequence served as targets. The model was trained to reconstruct the full diurnal cycle from sparse observations while also capturing temporal variability across training days.

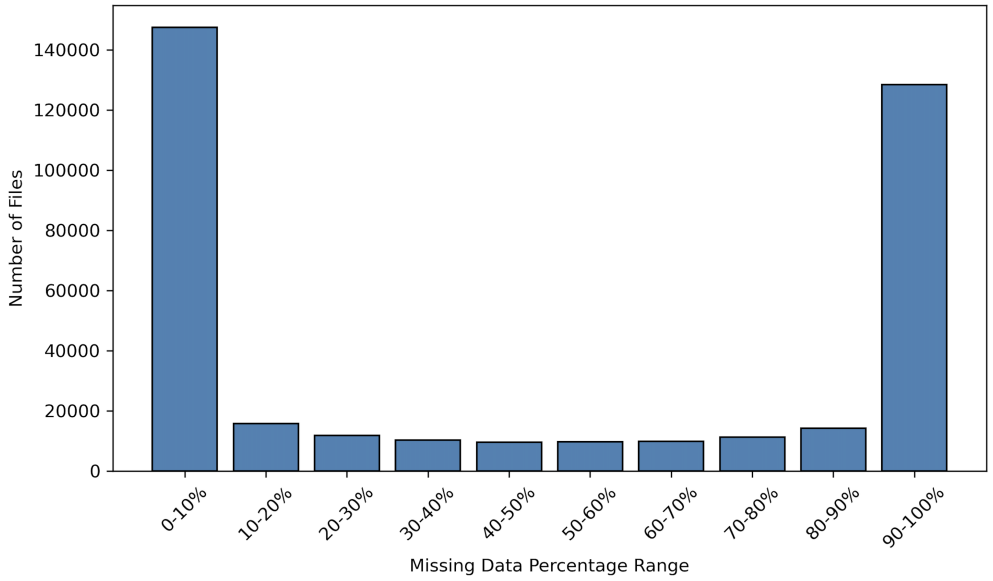

Figure 3.1: Average fraction of pixels with missing data in the preprocessed GOES-R satellite ABI Land Surface Temperature images of the year 2022 across the entire CONUS

Despite the advantage of high temporal resolution, GOES-R imagery suffers from extensive data loss due to environmental factors, cloud occlusions, and scan patterns. As depicted in Figure 3.1, the images contains significant number of missing pixels. Over 120,000 images contain 90-100% of missing values, which poses extreme challenge to use for model development. To address this, we constructed training and testing datasets from GOES-R observations over CONUS for 2022, excluding tiles in which missing values exceeded 95%. Data from 2023 were reserved exclusively for testing to assess generalization to unseen time periods, and in-situ weather station measurements were used as additional ground truth for evaluation.

The GOES-R temperature product is distributed in NetCDF format with geostationary projection metadata. To align these data with ancillary inputs, we converted them into georeferenced GeoTIFFs containing pixel-level coordinates. The transformation involved converting from the native geostationary projection to WGS-84 [41] using GDAL, with parameters such as sweep angle, inverse flattening, and semi-major axis derived from the metadata. This ensured spatial accuracy and consistency across all datasets used in the study.

Ancillary inputs included Köppen climate zones and elevation, both of which strongly influence diurnal patterns of land surface temperature. Climate zones affect the magnitude of heating and cooling cycles. For example, arid desert climates (BWh) often display large diurnal ranges due to clear skies, low humidity, and sparse vegetation, while humid tropical climates (Af) exhibit smaller ranges as dense vegetation and high humidity dampen fluctuations [42]. Elevation also plays an important role: higher elevations generally exhibit lower mean temperatures and more rapid nighttime cooling due to reduced atmospheric density, often producing larger diurnal ranges than adjacent lowland areas [43]. These factors together shape the timing and amplitude of diurnal cycles.

The DAYVIEW model was trained using input features composed of a randomly selected hourly image from GOES-R, together with the Köppen climate zone and elevation at the corresponding location, all spatially aligned. The target output consisted of the full sequence of 24 hourly images for the same day, representing the complete diurnal cycle.

3.2 Model Architecture

The goal of the DAYVIEW architecture is to generate a complete sequence of hourly land surface temperature maps from a single observed map. To accomplish this, we extend the Vision Transformer (ViT) [1](originally developed for computer vision tasks) by introducing three key modifications: (1) a regression head, (2) spatiotemporal encoding, and (3) a multi-modal data fusion strategy. Together, these allow the model to reconstruct diurnal cycles from sparse inputs

while integrating ancillary information. The resulting framework consists of two principal components: ViT-based encoder blocks and a set of regression layers that produce the 24 hourly maps.

The encoder blocks are initialized with weights from a pretrained ViT model [2]. Unlike the original design, which contains decoder blocks to reconstruct images, DAYVIEW removes the decoder and extends the final encoder block to perform regression tasks. This design directs the model's capacity toward sequence prediction rather than image reconstruction alone.

Training is carried out using the Masked Autoencoder (MAE) strategy [2], a form of self-supervised learning in which portions of the input are intentionally hidden. The model learns to infer the missing information from the visible data, thereby developing a deeper understanding of the underlying spatial structure and context. This approach enhances the encoder's ability to capture patterns in incomplete or partially available observations; a property essential for handling the high rates of missing data in satellite imagery.

As depicted in Figure 3.2, the input is represented as a three-channel image that is partitioned into 8×8 patches. This contrasts with the original ViT architecture, which divides a 224×224 image into 16×16 patches. During both training and inference, the encoder processes only the visible patches. To reduce the influence of tiles with limited valid data, patches with higher proportions of missing pixels are preferentially masked. This allows the encoder to learn representations that emphasize global spatial structure while remaining robust to missing values. 75% of patches are masked, which encourages the model to generalize from partial inputs and improves its ability to reconstruct the full diurnal sequence.

By combining masked self-supervised learning with spatiotemporal and multi-modal extensions, this architecture enables DAYVIEW to operate effectively at the CONUS scale, making it well suited for voluminous spatiotemporal data.

3.3 DAYVIEW Encoder Blocks with Regression Head

The encoder blocks in DAYVIEW serve two primary purposes. They fuse information from multiple datasets into a single coherent representation, and they produce concise embeddings that

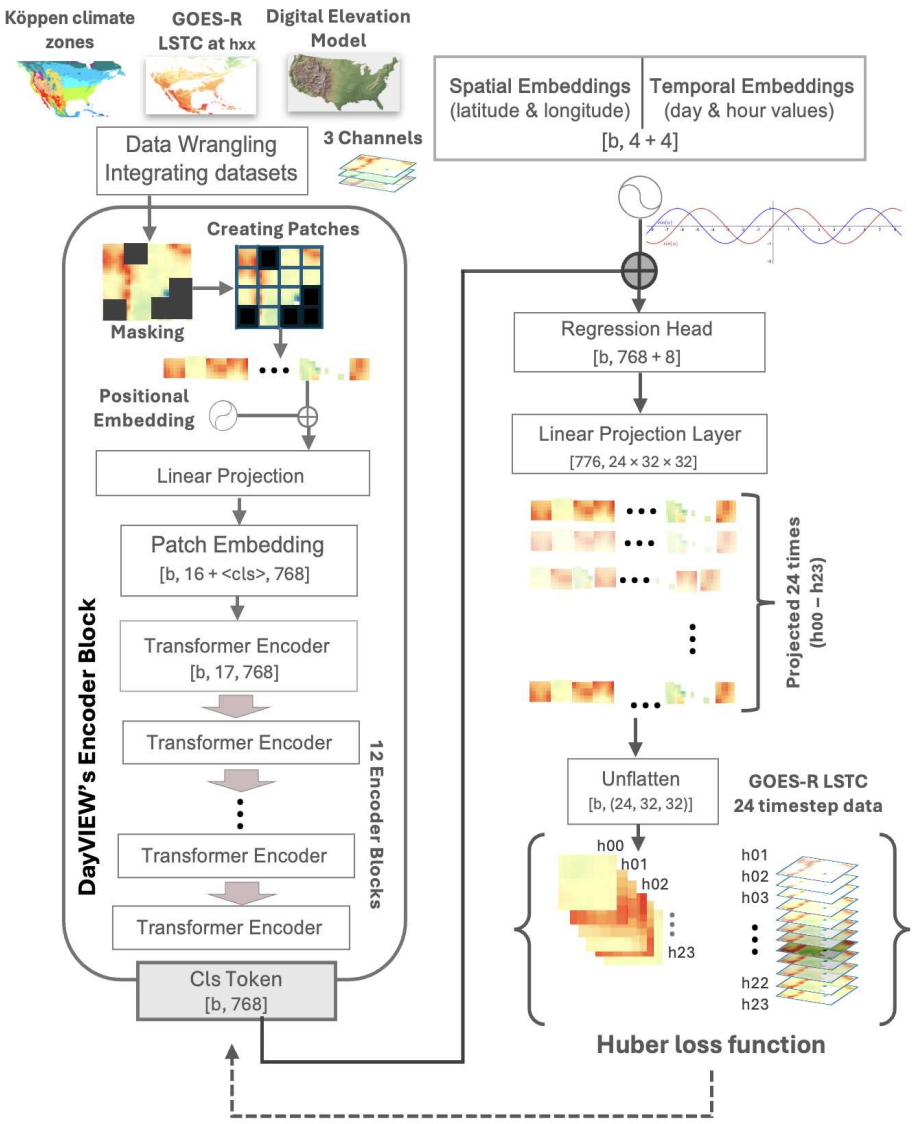

Figure 3.2: Overview of the DAYVIEW Framework for Generating Hourly Land Surface Temperature(LST) Map. The DAYVIEW framework consists of two key model components. The encoder blocks focuses on integrating GOES-R image and ancillary conditions and generating a representative embedding of the combined data. The regression multihead generates 24 hourly LST images without missing values over the target region.

retain the capacity to reconstruct missing values. To accomplish these objectives, DAYVIEW employs 12 stacked Vision Transformer (ViT) encoder blocks, drawing on the ViT's strength in capturing both local and global dependencies across spatial domains.

The process begins with the fusion of multi-modal inputs. Three datasets are aligned by their geospatial locations and combined into a single input image with three channels, each of size 32×32 pixels. Each pixel corresponds to an area of roughly 2km $\times$ 2km. These multi-channel images then serve as the input to the model. Within the encoder, each image is divided into patches of 8×8 pixels, producing 16 patches per image. Each patch is flattened into a one-dimensional vector and projected into a latent space of 768 hidden dimensions. The model preserves knowledge from the pretrained MAE-ViT by freezing some layers while fine-tuning the remainder on satellite data, thus balancing generalizable representations with task-specific adaptation.

The encoder blocks are initialized with weights from the pretrained MAE-ViT model, but only the encoder is retained; the decoder is excluded when composing the DAYVIEW framework. Whereas the original ViT decoder was designed to reconstruct full images, DAYVIEW redirects this capacity toward generating 24 sequential land surface temperature maps through a regression head. The encoder first produces a global embedding from the patch representations. This embedding, referred to as the global class token, summarizes the input image in a form that reflects relationships both within the observed patches and across the 23 unobserved hourly maps. Once trained, the global class token provides a representation that links the input observation to the full diurnal sequence.

To generate predictions, the global class token is concatenated with auxiliary embeddings that encode spatial and temporal context, as described in section 3.4

$$\mathbf{z}_{\text{fused}} = \Big[z_{\text{GCT}} \;\Big\|\; E_{\text{spatial}} \;\Big\|\; E_{\text{temporal}} \Big] \quad \in \mathbb{R}^{d_{\text{GCT}}+d_{\text{spatial}}+d_{\text{temporal}}} \tag{3.1}$$

where $\|$ denotes vector concatenation and z_{GCT} the global class token. This fused vector, z_{fused} is then passed through a fully connected layer and mapped to an output of size $24 \times 32 \times 32$. The output is reshaped into 24 frames, each corresponding to one hour of the diurnal cycle.

$$\hat{Y}_t = \text{Unflatten}\Big(f_{\text{reg}}(\mathbf{z}_{\text{fused}})\Big) \quad \in \mathbb{R}^{24\times32\times32} \tag{3.2}$$

where $\hat{Y}_t$ denotes the predicted map for hour t of the day.

Training this regression head requires a loss function robust to the extensive missing values in the input data. For this reason, we adopt a masked Huber loss [44]. Invalid pixels are excluded from the calculation, and the loss is computed only over valid observations. The Huber formulation applies mean squared error to small deviations and mean absolute error to larger deviations, thereby offering stability against outliers while preserving sensitivity to fine-scale differences:

$$L_\delta(p,t) = \begin{cases} \frac{1}{2}(p-t)^2, & \text{if } |p-t| \le \delta, \\ \delta\left(|p-t| - \frac{1}{2}\delta\right), & \text{if } |p-t| > \delta, \end{cases} \tag{3.3}$$

$$\begin{aligned} p &= \text{predicted LST value}, \\ t &= \text{ground-truth LST value}, \\ \delta &= \text{threshold separating quadratic and linear error} \end{aligned}$$

By integrating multi-modal fusion, encoder blocks, and a regression head guided by a robust loss function, the DAYVIEW framework extends the ViT–MAE paradigm to predict complete diurnal cycles over CONUS-scale domains.

3.4 Spatial and Temporal Encoding

The accurate modeling of diurnal land surface temperature requires explicit attention to both spatial and temporal context. From a spatial perspective: land cover, vegetation, and surface properties determine how heat is absorbed and released. Forests, croplands, urban areas, and bare soil, for example, display distinct thermal behaviors. Elevation, slope, and aspect influence the amount of solar radiation that reaches a given surface, while nearby water bodies further moderate

temperature through evaporation and thermal inertia. Regional climate provides the background conditions within which all of these local factors operate.

Temporal dynamics are equally important. Seasonal shifts in the angle of the sun alter the rate of surface heating, while soil and vegetation introduce lag effects by storing heat during the day and releasing it later. Ignoring these temporal signals would cause the model to miss the trends and transitions that define diurnal and seasonal variability. For this reason, DAYVIEW incorporates both geospatial and temporal features into its prediction of hourly temperature maps.

Geospatial information is provided through the latitude and longitude of the center of each input image. To ensure consistent representation across the globe, these coordinates are normalized and transformed into sine and cosine functions:

$$\big[\sin(\pi \cdot \text{lat}),\, \cos(\pi \cdot \text{lat}),\, \sin(\pi \cdot \text{lon}),\, \cos(\pi \cdot \text{lon})\big] W_s \tag{3.4}$$

where W_s is a learnable linear projection. This transformation preserves relative distances and spatial structure, enabling the network to recognize geographic relationships while maintaining inputs within a stable numerical range. The encoded spatial features are then passed through a projection layer that incorporates them into the model's embedding space.

Temporal encoding is based on the observation time recorded in the GOES-R metadata. Both the hour of the day and the day of the year are converted into cyclical representations using sine and cosine functions:

$$\left[\, \sin\!\big(2\pi \cdot \tfrac{\text{hour}}{24}\big)\,,\, \cos\!\big(2\pi \cdot \tfrac{\text{hour}}{24}\big)\,,\, \sin\!\big(2\pi \cdot \tfrac{\text{dayValue}}{365}\big)\,,\, \cos\!\big(2\pi \cdot \tfrac{\text{dayValue}}{365}\big) \,\right] W_t \tag{3.5}$$

where W_t is a learnable projection matrix. This representation eliminates artificial discontinuities in raw values. For instance, December 31 and January 1 appear numerically distant but are in fact consecutive days; the cyclical encoding preserves their proximity in the feature space. In this way, periodicity is represented geometrically: nearby hours and days map to vectors that lie close together, supporting effective interpolation and generalization. As with spatial encoding,

the sine–cosine transformation normalizes values to the range [–1, 1], which contributes to stable training.

The spatial and temporal vectors are then concatenated with the embeddings produced by the encoder blocks and passed to the regression head. The regression head maps this fused representation to a 24-frame output, with each frame corresponding to an hourly temperature map. By integrating spatiotemporal context in this manner, the architecture conditions its predictions on both local geography and temporal cycles, thereby improving its ability to capture daily temperature dynamics across diverse climates and landscapes.

By embedding both spatial and temporal context into its learning process, DAYVIEW strengthens its capacity to generalize across regions and seasons, ensuring that its diurnal reconstructions remain accurate at continental scale.

Table 3.1: Datasets used for the DAYVIEW Model Training and Evaluation

Dataset	Temporal Coverage	Size (GB)
GOES-R ABI product	1/2022-12/2022	78G
GOES-R preprocessed	1/2022-12/2022	14.4GB
Köppen Climactic	2022	532K
Digital Elevation Model	2020	706M

Chapter 4

Performance Benchmarks & Discussion

4.1 Dataset and Study Area

For this study, we focused on the entire CONUS (Continental United States) region. The hourly LST observations for the target dataset were retrieved from the GOES-R satellite ABI LST product, covering the period of the year 2022. The GOES-R ABI LST product is generated from ABI bands 14 and 15 using the split-window technique [36]. Data are available for each land or inland water pixel under *clear*, *probably clear*, or *probably cloudy conditions*. Consequently, a significant number of pixels in the product are unavailable, primarily due to cloud occlusion.

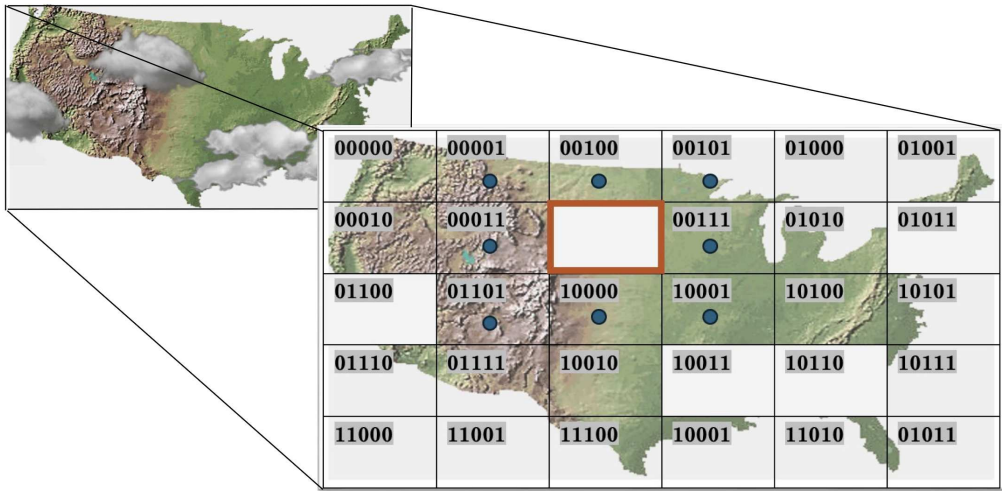

Figure 4.1: Learning geospatial proximity of data missing tiles due to cloud coverage with the help of neighboring tiles by assigning quad-hash

After preprocessing, which included filtering based on the availability of pixels. Images with more than 95% of missing data in the pre-processed, cropped regions based on the quad hash values were discarded. Images with lower missing values percentage were retained keeping in mind that their locations might contain valuable geographic information relevant for analysis. The presence of missing values also provided an opportunity to mask these regions and evaluate our

ability to predict the missing data. To capture relevant ancillary conditions, we integrated the Köppen Climatic Zone dataset [42] at 1 km resolution and the Digital Elevation Model at 30 m for elevation information. Table 3.1 summarizes the datasets used for DAYVIEW. We have selected bands 14 and 15 and filtered out images with high cloud coverage (more than 95% pixels with invalid values). Due to thetransition to GOES-18 and smaller datasest available for 2022, we used data collected in 2023. A total of 30GB of LST images were used to train and test DAYVIEW. These datasets were spatiotemporally aligned and split into an 80–20 ratio for training and testing in the modeling process.

4.2 Implementation Details

We initialized the weights in our encoder blocks using the pre-trained weights from ViT-MAE [1]. This library supports a wide variety of pretrained weights and non-pertained models specifically for image based tasks.

The input of the model contained hourly Land Surface skin images with Köppen climatic zones and elevation information attached as ancillary data points. Since the satellite provides entire continental U.S. as a single image, the data was partitioned into quad-hash tiles for every hour where each tile was resized into 32×32 pixels. As depicted in Figure 4.1, a quad-hash recursively and hierarchically divides a two-dimensional space into four quadrants. All input images were cropped and aligned using quad-hash. These input data contains channels of LST, Köppen climactic zone information and elevation, and the values were normalized independently using mean and standard deviation for each channel. A Masked Transformer encoder (ViT-B/16) model was implemented to learn the spatial and temporal patterns in hourly GOES-R imagery [1]. Within a quad-hash tile, patches with higher proportion of pixels containing invalid or missing values were used for masking.

The model was trained with 80% of input data with all three channels. We trained the model using the AdamW optimizer with an initial learning rate of 0.0001, for over 100 epoch with a batch size of 64. For evaluation and fine-tuning the model, Peak Signal-to-Noise Ratio(PSNR) and

Huber loss function were used. During calculating of training loss, missing values in the target were ignored. The ViT model included Gaussian Error Linear Unit(GELU) activation functions in its transformer blocks. The model outputs a 24 band, single channel representing predicted LST values for a full diurnal cycle, given a single input at a specific hour of that day. DAYVIEW was run on a single NVIDIA A100 GPU which is 40GB, with a maximum run time of 200 hours.

4.3 Model Accuracy and Ablation Studies

Table 4.1: Evaluation Metrics for Image Predicting Models. ↑ indicates higher is better, ↓ indicates lower is better. The best performance is highlighted in red, and the second-highest performance is shown in blue.

Comparison	**MAE** ↓	**RMSE** ↓	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	R^2 ↑
ViT	0.0326	0.0378	29.6850	0.3960	0.4088	0.8324
ViT-MAE	0.0343	0.0395	29.1742	0.3935	0.4060	0.8223
Temporal-Sequence like model	0.0741	0.0754	26.0121	0.2567	0.4407	0.1863
TimeSformer like model	0.0746	0.0759	32.7172	0.3997	0.4099	0.1818
GRU	0.0297	0.0317	31.0985	0.4014	0.3812	0.8445
Bi-LSTM	0.0276	0.0299	29.1603	0.3980	0.4051	0.9020
Ablation						
ViTMAE (no ancillary data)	0.0331	0.0355	31.2839	0.3998	0.4019	0.8345
ViTMAE (temp. embedding)	0.0378	0.0395	28.5869	0.4029	0.3980	0.8018
ViTMAE (spat. embedding)	0.0385	0.0344	30.3279	0.3597	0.4007	0.7775
DAYVIEW (Our Model)	**0.0290**	**0.0266**	**32.7955**	**0.4052**	**0.3980**	**0.9091**

Evaluation metrics

To evaluate our model performance at the pixel level, we have used standard error metrics for regression models such as Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). We report the coefficient of determination(R^2), that tells us the percentage of variation in the predicted images from the model and provides a global indication of performance across all valid pixels indicating how well the model fits the data [45]. To evaluation the accuracy of generative images, we used peak signal-to-noise ration (PSNR) [46] and Structural Similarity index measure(SSIM) [47], which evaluate both fidelity and structural similarity between reconstructed and true LST fields.

Additionally, we included Learned Perceptual Image Patch Similarity(LPIPS) [48] to measure perceptual similarity from human-vision perspective, all of which provides a complementary assessment if spatial and temporal correctness in the predicted diurnal cycles. Collectively, these metrics allow a comprehensive evaluation of both fine-grained accuracy and overall structural consistence of the model outputs.

Model Accuracy As shown in Table 4.1, DAYVIEW achieves an MAE of 0.0266 and an RMSE of 0.0290, outperforming all baseline models. These low error values indicate that the model can accurately reconstruct land surface temperature values at pixel level, even from a single observation. It also demonstrates superior image reconstruction quality with a PSNR of 32.80 dB and an SSIM of 0.4052 confirming that the model preserves fine-scale spatial details in the reconstructed images, minimizing noise to the ground truth indicated by PSNR values. In addition to the low errors, DAYVIEW acquires a high coefficient of determination(0.9091), indicating that over 90% of the variance in the ground-truth diurnal LST values which shows that there is a strong agreement between the overall temporal and spatial variability of the reconstructed temperature maps. The SSIM score indicates large spatial and structural patterns are being captured in the diurnal temperature variation which is equally important for maintaining the physical consistency of the the reconstructed thermal maps. DAYVIEW maintains perceptual similarity with an LPIPS score of 0.3980 indicating that the predicted LST sequences are perceptually consistent with the ground truth. This proves that the model was able to reproduce patterns that are meaningful at the human perception scale. All of these evaluations where the low values of MAE and RMSE, strong R^2 score along with high PSNR and moderate SSIM and LPIPS suggest that DAYVIEW balances pixel-wise precision, structural fidelity and perceptual patterns for diurnal variations of land surface temperature data.

For comparison, we also evaluated several baseline models including: a standard ViT, ViT-MAE, a transformer variants inspired by TimeSformer [49], Temporal-sequential convolution Transformer [50] (capturing temporal dependencies using a single timestamp information), and recurrent architectures such as GRU and Bi-LSTM. All were trained on the same three–channel dataset ex-

cept for ViTMAE where no ancillary information was provided. Among them, Bi-LSTM achieved the closest MAE(0.0276) indicating the models performance in calculating pixel-level predictions, but both GRU and Bi-LSTM underperformed in PSNR and SSIM which meant while they capture temporal trends well, they fail to preserve the fine spatial detail inherent in the imagery.

The ViT-based architecture of DAYVIEW efficiently integrates spatial and temporal context which allows it to scale more efficiently than recurrent models, making it better suited for handling the high data volumes and continental coverage required in voluminous spatial observational data settings while enabling reconstruction from a single observation. These results directly support **RQ-1**, demonstrating that a single observation, when processed through the DAYVIEW architecture, is sufficient to reconstruct accurate diurnal profiles at both pixel-level precision and structural level of fidelity that recurrent and baseline transformer models cannot achieve, making it an optimum model for practical applications such as urban heat monitoring and agricultural management.

Ablation studies An ablation study systematically removes or isolates components of the model to reveal their individual contributions. In our context, this allows us to assess how temporal and spatial embeddings each affect the reconstruction of diurnal LST cycles. To assess the individual contributions of the learnable temporal and spatial embeddings, we began with the base model and then added each component in turn. Introducing only temporal information (ViTMAE + temporal embedding) increased SSIM from 0.3960 to 0.4029, indicating that temporal context helps the model capture the smoothness of diurnal transitions and improves the variance of reconstructed temperature dynamics as reflected by higher R^2 values . By contrast, adding only spatial information (ViTMAE + spatial embedding) reduced RMSE from 0.0378 to 0.0344, showing that geographic cues provide key context for local variations, improving the stability and enhance the stability of pixel–wise predictions but could notice the decrease in determination coefficient indicating less variance in the diurnal cycle values predicted. However, while each embedding improved performance on its own; neither matched the accuracy of the full DAYVIEW model. The combined inclusion of temporal and spatial information proved necessary to achieve the highest performance, confirming their joint importance for robust diurnal reconstruction. These results directly support

RQ-2, demonstrating that ancillary data, when effectively integrated, provide essential context for modeling the full dynamics of diurnal LST cycles.

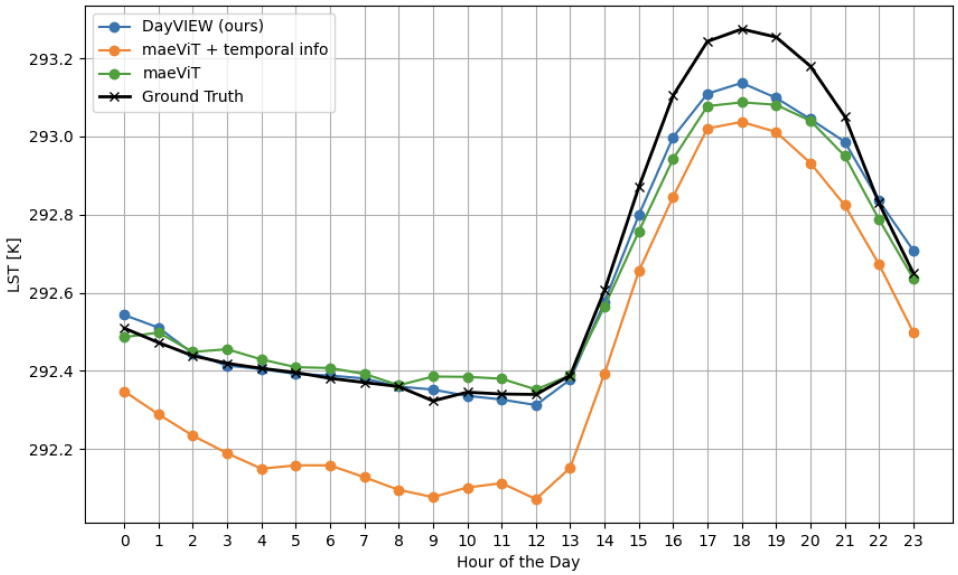

Figure 4.2: Hourly LST predictions of models and the ground truth

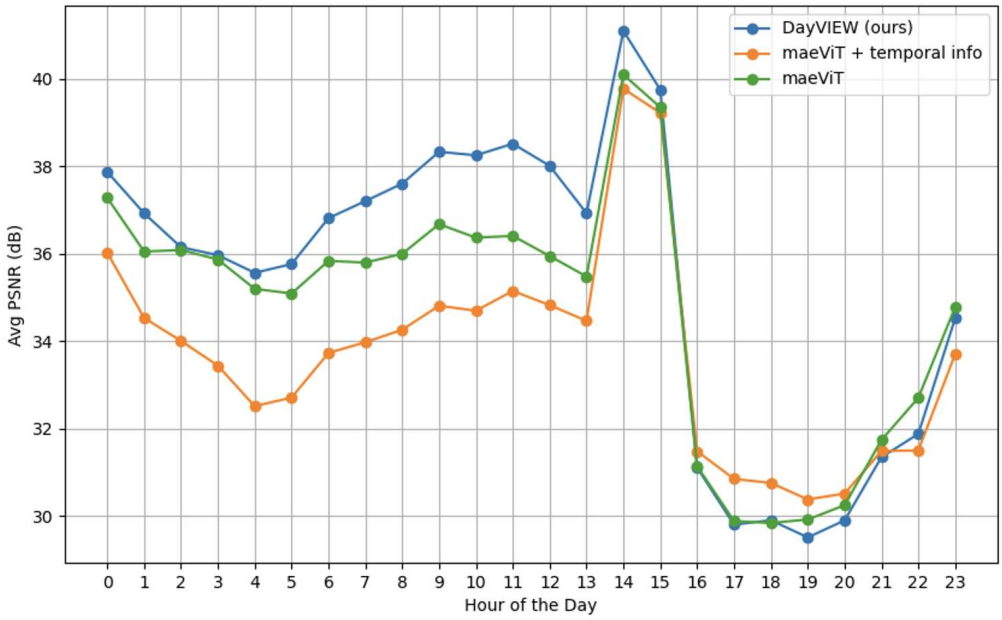

Figure 4.3: Hourly PSNR values calculated on model predictions

Figure 4.2 and Figure 4.3 depict the hourly LST predictions and PSNR values from both our model and the baseline models, alongside the ground truth provided by the GOES-R dataset. Among these comparisons, DAYVIEW shows the closest alignment to the observed diurnal curve, where it consistently outperforms all baselines. The basic ViTMAE model performs well during the first portion of the diurnal cycle but fails to maintain fine-grained temporal accuracy during the later hours. In contrast, DAYVIEW sustains performance across the full cycle and achieves the highest average PSNR when results are aggregated across all hours of the day. These results corroborate **RQ-1**, demonstrating that a single sparse observation, when processed through the DAYVIEW, is sufficient to reconstruct the diurnal sequence with fidelity.

4.4 Model Sensitivity Analysis

We performed a set of model sensitivity analyses across different land cover types using PSNR as the evaluation metric. The test data were predicted and compared to the ground truth to assess reconstruction quality and examine how it varied with underlying land cover characteristics. Table 4.2 summarizes the results of the model sensitivity analysis.

Table 4.2: Model sensitivity across Land Cover Types

Land Cover Type	PSNR
Developed, Open Space	**37.31**
Developed, Low Intensity	**38.02**
Developed, Medium Intensity	**37.62**
Developed, High Intensity	36.86
Barren Land (Rock/Sand/Clay)	34.60
Deciduous Forest	35.86
Evergreen Forest	36.19
Mixed Forest	27.13
Shrub/Scrub	34.87
Grassland/Herbaceous	35.14
Pasture/Hay	32.77
Cultivated Crops	27.31
Woody Wetlands	33.86

All developed area classes among the land cover types achieved the highest PSNR values, with a maximum of 38, indicating strong fidelity in these regions. This suggests that the model effectively captured the LST diurnal pattern over structural and man-made environments. Forest classes, such as deciduous and evergreen forests, also achieved relatively high scores, though slightly lower than developed areas. The drop in accuracy for mixed forests likely stems from the complexity of diverse natural canopy structures compared to the more uniform buildings and structures in developed areas.

In contrast, cultivated crops and pasture classes showed lower PSNR values, likely due to frequent seasonal variations. Different stages of crop development alter canopy height and coverage throughout the season, while crop rotation introduces interannual variability, both of which challenge accurate diurnal pattern estimation. Barren lands, being more uniform and less textured, were easier to predict than patchy or dynamic landscapes, as demonstrated by the model's performance.

Overall, the model demonstrated consistently high accuracy in areas with human-made structures, such as developed regions, but struggled more in heterogeneous or highly variable land cover types like forests and croplands.

Chapter 5

Conclusions & Future Work

Our methodology targets the the reconstruction of diurnal land surface temperature (LST) cycles at scale and our design was informed by three interrelated research questions. Alongside our methodology, we have performed benchmarks, ablation experiments, and sensitivity analyses; these allow use to frame our conclusions in terms of these questions.

RQ1 explored whether a data–driven model could estimate complete diurnal cycles for LST from sparse temporal measurements. The results demonstrate that a single LST observation, regardless of its acquisition time, can serve as a sufficient basis for reconstructing the twenty–four hourly profiles of a day. DAYVIEW consistently achieved lower errors and stronger structural similarity than competing models, confirming that our generative sequence approach is an effective solution to the sparsity problem.

RQ2 explored whether the integration of ancillary data improves the capacity to capture diurnal fluctuations. The ablation studies show that incorporating climatic zones and elevation significantly enhances prediction quality, particularly in reducing error at the pixel level and capturing smoothness across hourly transitions. Temporal embeddings alone increased structural similarity, while spatial embeddings improved stability of pixel–wise predictions. Together, their combination produced the highest accuracy. In particular, environmental context is essential to robust diurnal reconstruction.

RQ3 explored whether the model could scale to large spatial extents while maintaining generalizability. The continental–scale experiments across the CONUS region validates that DAYVIEW preserves accuracy and robustness across diverse land cover types and climatic regimes. Sensitivity analyses indicate that the model adapts well to variability in geography and seasonality; this underscores the suitability DAYVIEW for large geographical extents.

Taken together, these findings demonstrate the suitability of DAYVIEW as a framework capable of reconstructing diurnal LST cycles from observations that are both limited and imperfect. The

model's capacity to integrate sparse measurements with environmental context, and to extend its predictions reliably across continental scales, underscores its value as a potential tool for large–area environmental monitoring.

As part of future work we will pursue two directions. The first will examine the translational impact of this methodology on other diurnal processes (such as solar–induced fluorescence and evapotranspiration) where the need for complete hourly profiles is equally pressing. The second avenue will investigate how DAYVIEW can be adapted for continuous inferencing at targeted spatial extents, enabling its integration into decision support systems for applications in agriculture, climate adaptation, and environmental management.

Bibliography

[1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[3] Stephan Bojinski, Michel Verstraete, Thomas C Peterson, Carolin Richter, Adrian Simmons, and Michael Zemp. The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, 95(9):1431–1443, 2014.

[4] MJ Salinger and GM Griffiths. Trends in new zealand daily temperature and rainfall extremes. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(12):1437–1452, 2001.

[5] DC Reicosky, LJ Winkelman, JM Baker, and DG Baker. Accuracy of hourly air temperatures calculated from daily minima and maxima. *Agricultural and forest Meteorology*, 46(3):193–209, 1989.

[6] F.M. Göttsche and F.S. Olesen. Modelling diurnal cycles of land surface temperature using msg seviri data. *Remote Sensing of Environment*, 113:2304–2316, 2009.

[7] Menglin Jin and Robert E Dickinson. Interpolation of surface radiative temperature measured from polar orbiting satellites to a diurnal cycle: 1. without clouds. *Journal of Geophysical Research: Atmospheres*, 104(D2):2105–2116, 1999.

[8] Rachel T Pinker, Donglian Sun, Meng-Pai Hung, Chuan Li, and Jeffrey B Basara. Evaluation of satellite estimates of land surface temperature from goes over the united states. *Journal of Applied Meteorology and Climatology*, 48(1):167–180, 2009.

[9] Anand K Inamdar, Andrew French, Simon Hook, Greg Vaughan, and William Luckett. Land surface temperature retrieval at high spatial and temporal resolutions over the southwestern united states. *Journal of Geophysical Research: Atmospheres*, 113(D7), 2008.

[10] Kate Duffy, Thomas J Vandal, and Ramakrishna R Nemani. Multisensor machine learning to retrieve high spatiotemporal resolution land surface temperature. *IEEE Access*, 10:89221–89231, 2022.

[11] Shengjie Liu, Siqin Wang, and Lu Zhang. Daily land surface temperature reconstruction in landsat cross-track areas using deep ensemble learning with uncertainty quantification. *arXiv preprint arXiv:2502.14433*, 2025.

[12] Qin Su, Yuan Yao, Cheng Chen, and Bo Chen. Generating a 30 m hourly land surface temperatures based on spatial fusion model and machine learning algorithm. *Sensors*, 24(23):7424, 2024.

[13] Katja Kustura, David Conti, Matthias Sammer, and Michael Riffler. Harnessing multi-source data and deep learning for high-resolution land surface temperature gap-filling supporting climate change adaptation activities. *Remote Sensing*, 17(2):318, 2025.

[14] J. Han et al. A time-continuous land surface temperature (lst) data product using deep learning and multi-source observations. *Science of the Total Environment*, 2024.

[15] Edwin Goh, Alice R Yepremyan, Jinbo Wang, and Brian Wilson. Maesstro: Masked autoencoders for sea surface temperature reconstruction under occlusion. *EGUsphere*, 2023:1–20, 2023.

[16] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[17] Hugo Chan-To-Hing and Bharadwaj Veeravalli. Fus-mae: A cross-attention-based data fusion approach for masked autoencoders in remote sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 6953–6958. IEEE, 2024.

[18] Zefeng Wu, Hongfen Teng, Haoxiang Chen, Lingyu Han, and Liangliang Chen. Reconstruction of gap-free land surface temperature at a 100 m spatial resolution from multidimensional data: A case in wuhan, china. *Sensors*, 23(2):913, 2023.

[19] Teerapong Panboonyuen, Chaiyut Charoenphon, and Chalermchon Satirapod. Mevit: a medium-resolution vision transformer for semantic segmentation on landsat satellite imagery for agriculture in thailand. *Remote Sensing*, 15(21):5124, 2023.

[20] Mohammadreza Heidarianbaei, Hubert Kanyamahanga, and Mareike Dorozynski. Temporal vit-u-net tandem model: Enhancing multi-sensor land cover classification through transformer-based utilization of satellite image time series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10:169–177, 2024.

[21] Nawel Slimani, Imen Jdey, and Monji Kherallah. Improvement of satellite image classification using attention-based vision transformer. In *ICAART (3)*, pages 80–87, 2024.

[22] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.

[23] Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. Knowledge-guided machine learning: Current trends and future prospects. *arXiv preprint*, arXiv:2403.15989, 2024.

[24] Paahuni Khandelwal, Sangmi Lee Pallickara, and Shrideep Pallickara. Deepsoil: A science-guided framework for generating high precision soil moisture maps by reconciling measurement profiles across in-situ and remote sensing data. In *Proceedings of the 32nd ACM In-*

ternational Conference on Advances in Geographic Information Systems, pages 233–246, 2024.

[25] Paahuni Khandelwal, Jeffrey D Niemann, David J Mulla, Shrideep Pallickara, and Sangmi Lee Pallickara. Subterra: Estimating soil moisture at root zone depths using science-guided learning. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 328–335. IEEE, 2025.

[26] Kevin Bruhwiler, Paahuni Khandelwal, Daniel Rammer, Samuel Armstrong, Sangmi Lee Pallickara, and Shrideep Pallickara. Lightweight, embeddings based storage and model construction over satellite data collections. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 246–255. IEEE, 2020.

[27] Samuel Armstrong, Paahuni Khandelwal, Dhruv Padalia, Gabriel Senay, Darin Schulte, Allan Andales, F Jay Breidt, Shrideep Pallickara, and Sangmi Lee Pallickara. Attention-based convolutional capsules for evapotranspiration estimation at scale. *Environmental Modelling & Software*, 152:105366, 2022.

[28] Abdul Matin, Samuel Armstrong, Saptashwa Mitra, Shrideep Pallickara, and Sangmi Lee Pallickara. Rapid betweenness centrality estimates for transportation networks using capsule networks. In *2022 Fourth International Conference on Transdisciplinary AI (TransAI)*, pages 89–96. IEEE, 2022.

[29] Paahuni Khandelwal, Samuel Armstrong, Abdul Matin, Shrideep Pallickara, and Sangmi Lee Pallickara. Cloudnet: A deep learning approach for mitigating occlusions in landsat-8 imagery using data coalescence. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 117–127. IEEE, 2022.

[30] Saptashwa Mitra, Daniel Rammer, Shrideep Pallickara, and Sangmi Lee Pallickara. Glance: A generative approach to interactive visualization of voluminous satellite imagery. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 359–367. IEEE, 2021.

[31] Tanjim Bin Faruk, Abdul Matin, Shrideep Pallickara, and Sangmi Lee Pallickara. Accounting for spatial variability with the histogram of oriented gradients based masking improves performance of masked autoencoder over hyperspectral satellite imagery (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29365–29367, 2025.

[32] Abdul Matin, Tanjim Bin Faruk, Shrideep Pallickara, and Sangmi Lee Pallickara. Hyperkd: Distilling cross-spectral knowledge in masked autoencoders via inverse domain shift with spatial-aware masking and specialized loss. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–11, 2025.

[33] Andrei Bachinin, Rupasree Dey, Paahuni Khandelwal, Sam Leuthold, M Francesca Cotrufo, Shrideep Pallickara, and Sangmi Lee Pallickara. Science-informed multitask transformer for soil property prediction from ftir spectroscopy. In *2025 IEEE International Conference on eScience (eScience)*, pages 48–57. IEEE, 2025.

[34] Rupasree Dey, Abdul Matin, Everett Lewark, Tanjim Bin Faruk, Andrei Bachinin, Sam Leuthold, M Francesca Cotrufo, Shrideep Pallickara, and Sangmi Lee Pallickara. Deepsalt: Bridging laboratory and satellite spectra through domain adaptation and knowledge distillation for large-scale soil salinity estimation. In *IEEE International Conference on Big Data 2025*, 2025.

[35] Abdul Matin, Paahuni Khandelwal, Shrideep Pallickara, and Sangmi Lee Pallickara. Discern: Leveraging knowledge distillation to generate high resolution soil moisture estimation from coarse satellite data. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1222–1229. IEEE, 2023.

[36] Timothy J Schmit, Mathew M Gunshor, W Paul Menzel, James J Gurka, Jun Li, and A Scott Bachmeier. Introducing the next-generation advanced baseline imager on goes-r. *Bulletin of the American Meteorological Society*, 86(8):1079–1096, 2005.

[37] Michael A Wulder, David P Roy, Volker C Radeloff, Thomas R Loveland, Martha C Anderson, David M Johnson, Sean Healey, Zhe Zhu, Theodore A Scambos, Nima Pahlevan, et al. Fifty years of landsat science and impacts. *Remote Sensing of Environment*, 280:113195, 2022.

[38] Carl F Schueler, Thomas F Lee, and Steven D Miller. Viirs constant spatial-resolution advantages. *International Journal of Remote Sensing*, 34(16):5761–5777, 2013.

[39] Timothy J Schmit, Paul Griffith, Mathew M Gunshor, Jaime M Daniels, Steven J Goodman, and William J Lebair. A closer look at the abi on the goes-r series. *Bulletin of the American Meteorological Society*, 98(4):681–698, 2017.

[40] M Platings and AM Day. Compression of large-scale terrain data for real-time visualization using a tiled quad tree. In *Computer Graphics Forum*, volume 23, pages 741–759. Wiley Online Library, 2004.

[41] B LOUIS Decker. World geodetic system 1984. 1986.

[42] Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. World map of the köppen-geiger climate classification updated. 2006.

[43] Peter L Guth, Adriaan Van Niekerk, Carlos H Grohmann, Jan-Peter Muller, Laurence Hawker, Igor V Florinsky, Dean Gesch, Hannes I Reuter, Virginia Herrera-Cruz, Serge Riazanoff, et al. Digital elevation models: Terminology and definitions. *Remote Sensing*, 13(18):3581, 2021.

[44] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[45] Dimo Brockhoff, Tobias Wagner, and Heike Trautmann. On the properties of the r2 indicator. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*, pages 465–472, 2012.

[46] Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pages 37–38. IEEE, 2012.

[47] Dominique Brunet, Edward R Vrscay, and Zhou Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011.

[48] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D Roads, Michael C Mozer, and Richard S Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE international conference on image processing (ICIP)*, pages 4277–4281. IEEE, 2017.

[49] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Icml*, volume 2, page 4, 2021.

[50] Lei Huang, Feng Mao, Kai Zhang, and Zhiheng Li. Spatial-temporal convolutional transformer network for multivariate time series forecasting. *Sensors*, 22(3):841, 2022.

Appendix A

License

Colorado State University LaTeX Thesis Template

by Elliott Forney – 2017

This is free and unencumbered software released into the public domain.

Anyone is free to copy, modify, publish, use, compile, sell, or distribute this software, either in source code form or as a compiled binary, for any purpose, commercial or non-commercial, and by any means.

In jurisdictions that recognize copyright laws, the author or authors of this software dedicate any and all copyright interest in the software to the public domain. We make this dedication for the benefit of the public at large and to the detriment of our heirs and successors. We intend this dedication to be an overt act of relinquishment in perpetuity of all present and future rights to this software under copyright law.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.